UDC 551.51:532.59:517.9

Numerical Solutions of the One-Dimensional Primitive Equations Using Galerkin Approximations With Localized Basis Functions

HSUAN-HENG WANG and PAUL HALPERN—*IBM Scientific Center, Palo Alto, Calif.*

JIM DOUGLAS, JR., and TODD DUPONT—*Mathematics Department, University of Chicago, Chicago, Ill.*

ABSTRACT—The Galerkin method is applied to a pair of linear and then nonlinear primitive (wave) equations. This results in a system of ordinary differential equations. Procedures are included for generating the coefficient matrices of the system of ordinary differential equations when piecewise Hermite cubic functions are used as basis functions. It is demonstrated that this system can be efficiently solved by an implicit method. Numerical examples show that integration using the Galerkin method is more efficient than the corresponding finite-difference method with central differences in space.

1. INTRODUCTION

Finite-difference techniques have long been the standard method of numerically solving meteorologically related boundary value and mixed initial and boundary value problems. A review of many commonly used schemes was given by Kurihara (1965) and later by Grammeltvedt (1969). On the other hand, numerical methods based on the variational formulation of the physical problem have emerged in recent years as strong competitors to the finite-difference methods. Especially, a renewed interest in the Galerkin (1915) approximation has been kindled by the establishment of rigorous error bounds for such approximations (Varga 1970, Price and Varga 1970, Swartz and Wendroff 1969) coupled with recent developments in spline functions (Ahlberg et al. 1967) and piecewise Hermite polynomials functions (Birkhoff et al. 1968). In a recent paper, Douglas and Dupont (1970) established error bounds for a number of numerical schemes when the Galerkin method is applied to a system of parabolic equations. These error bounds were established for both linear and nonlinear problems. In this paper, we apply the Galerkin method to a pair of first-order, quasi-linear hyperbolic wave equations.

In the first part of the paper, we present a detailed discussion of the Galerkin method and the means of implementation for the linear wave equations. The resulting system of ordinary differential equations is solved by the Crank-Nicolson implicit method (Kurihara 1965). The results are compared to the exact solution and to results using the corresponding implicit finite-difference scheme. The implementation of the boundary conditions for the Galerkin procedure will also be demonstrated. In the second part of the paper, the Galerkin procedure for the nonlinear wave equations is discussed. A numerical example is presented and compared to finite-difference solutions. The aim of this paper is to show the relative merits of the Galerkin method versus the finite-difference method in solving the problem of wave propagation when the same time differencing scheme is used in both computations.

2. PHYSICAL MODEL AND GOVERNING EQUATIONS

The physical system to be considered initially is one dimensional and composed of a single homogeneous layer of fluid that is in hydrostatic equilibrium. The set of governing differential equations is

$$\frac{\partial u'}{\partial t}+u'\frac{\partial u'}{\partial x}+g\frac{\partial h'}{\partial x}=0 \quad t\geq 0$$

and (1)

$$\frac{\partial h'}{\partial t}+h'\frac{\partial u'}{\partial x}+u'\frac{\partial h'}{\partial x}=0 \quad 0\leq x\leq L$$

where u' represents the fluid velocity in the x direction, h' represents the depth of the fluid, and g is the gravity constant.

If we employ the standard perturbation approximations and if we neglect the advection terms, eq (1) reduce to

$$\frac{\partial u}{\partial t}+g\frac{\partial h}{\partial x}=0$$

and (2)

$$\frac{\partial h}{\partial t}+H\frac{\partial u}{\partial x}=0$$

where H is the mean depth of the fluid and u and h are the perturbation quantities. To obtain a solution to eq (2), we must specify initial values, $u(x,0)$ and $h(x,0)$.

Proper boundary conditions for this set of equations require specification of either u or h or their derivatives

at $x=0$ and $x=L$. For example, the rigid wall condition is

$$u(0, t)=u(L, t)=0. \tag{3}$$

When using finite-difference methods for solving eq (2) with the boundary conditions in eq (3), one must take special care in evaluating h on the boundary. Moretti (1969) has recently discussed this problem. To avoid this inconvenience, one can employ the periodic condition. This takes the form

$$u(x, t)=u(x+L, t) \tag{4}$$

and

$$h(x, t)=h(x+L, t). \tag{5}$$

It is shown in the next section that no boundary problem arises when the Galerkin method is used. We discuss methods of numerically solving eq (2) in sections 3-5. In sections 6 and 7, numerical solutions to eq (1) are discussed.

3. THE GALERKIN METHOD

a. A Variational Formulation

We shall describe a variational form of eq (2). Let S be the space of all real-valued, piecewise, continuously differentiable functions, $v(x)$, on $[0, L]$. Multiply eq (2) by v and integrate from 0 to L with respect to x, obtaining

$$\int_0^L\left(\frac{\partial u}{\partial t}+g\frac{\partial h}{\partial x}\right)v dx=0$$

and (6)

$$\int_0^L\left(\frac{\partial h}{\partial t}+H\frac{\partial u}{\partial x}\right)v dx=0.$$

The above equations should be satisfied at each time $t>0$ and for any arbitrary function $v\epsilon S$. In addition, the initial condition should also be satisfied. (The implementation of boundary conditions will be discussed in subsection 3d.) This is a variational formulation of eq (2). The variational equations require that the solutions be subject to the same restriction as the test function v; that is, they belong to the space S. Any approximate solution to eq (6) will be an approximate solution to eq (2). A more detailed theoretical discussion on variational methods is given by Mikhlin (1964).

b. Galerkin Approximation

The Galerkin method approximates the solution to eq (6) and can be described as follows. First, choose a finite-dimensional subspace S_{2N} of S. Replacing $u(x,t)$ and $h(x,t)$ in eq (6) by $U(x,t)$ and $P(x,t)$, we obtain the Galerkin equations,

$$\int_0^L\left(\frac{\partial U}{\partial t}+g\frac{\partial P}{\partial x}\right)v dx=0$$

and (7)

$$\int_0^L\left(\frac{\partial P}{\partial t}+H\frac{\partial U}{\partial x}\right)v dx=0,$$

for $t\geq 0$ and all $v\epsilon S_{2N}$. Here, U and P are the Galerkin approximations to u and h for each time $t\geq 0$, and they are elements of S_{2N}.

Let $v_1(x)$, $v_2(x)$, . . . , $v_{2N}(x)$ be a basis for S_{2N}. Then, for any time $t>0$, $U(x,t)$ and $P(x,t)$ can be expressed as

$$U(x,t)=\sum_{i=1}^{2N}\alpha_i(t)v_i(x)$$

and (8)

$$P(x,t)=\sum_{i=1}^{2N}\beta_i(t)v_i(x),$$

respectively. Equations (7) will be satisfied for all $v\epsilon S_{2N}$ if they are satisfied for $v=v_1, v_2, \ldots, v_{2N}$. By substituting eq (8) into (7), these $4N$ relations are transformed into the following $4N$ equations:

$$\sum_{i=1}^{2N}\int_0^L\frac{d\alpha_i}{dt}v_iv_j dx+g\sum_{i=1}^{2N}\int_0^L\beta_i\frac{dv_i}{dx}v_j dx=0$$

and (9)

$$\sum_{i=1}^{2N}\int_0^L\frac{d\beta_i}{dt}v_iv_j dx+H\sum_{i=1}^{2N}\int_0^L\alpha_i\frac{dv_i}{dx}v_j dx=0$$

where $j=1, 2, \ldots, 2N$. We can write eq (9) in matrix notation as

$$\mathbf{C}\frac{d\boldsymbol{\alpha}}{dt}+g\mathbf{A}\boldsymbol{\beta}=0$$

and (10)

$$\mathbf{C}\frac{d\boldsymbol{\beta}}{dt}+H\mathbf{A}\boldsymbol{\alpha}=0$$

where $\mathbf{C}$ and $\mathbf{A}$ are $2N\times 2N$ matrices for which (i,j) elements, C_{ij} and A_{ij}. are, respectively,

$$C_{ij}=\int_0^L v_iv_j dx$$

and (11)

$$A_{ij}=\int_0^L\frac{dv_i}{dx}v_j dx,$$

and $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are $2N$-dimensional column vectors with transposes

$$\boldsymbol{\alpha}^T=(\alpha_1, \alpha_2, \ldots, \alpha_{2N})$$

and

$$\boldsymbol{\beta}^T=(\beta_1, \beta_2, \ldots, \beta_{2N}),$$

respectively.

We note that the Galerkin method has transformed the original partial differential equations into a system of ordinary differential equations [eq (10)] in the coefficients of the basis functions. For each partial differential equation, there is a system of $2N$ ordinary differential equations associated with $2N$ basis functions. Before we can solve eq (10), they must be complemented by the initial values $\boldsymbol{\alpha}(0)$ and $\boldsymbol{\beta}(0)$. These initial values can be obtained by projecting $u(x,0)$ and $h(x,0)$ into S_{2N}; that is, by solving the following equations:

$$\sum_{i=1}^{2N}\int_0^L\alpha_i(0)v_iv_j dx=\int_0^L u(x,0)v_j dx$$

and (12)

$$\sum_{i=1}^{2N}\int_0^L\beta_i(0)v_iv_j dx=\int_0^L h(x,0)v_j dx$$

where $j=1, 2, \ldots, 2N$, or in matrix notation

and

$$\mathbf{C}\boldsymbol{\alpha}(0)=\mathbf{q}$$
$$\mathbf{C}\boldsymbol{\beta}(0)=\mathbf{s} \tag{13}$$

where the jth component of $\mathbf{q}$ is $q_j=\int_0^L u(x,0)v_j dx$ and kth component of $\mathbf{s}$ is $s_k=\int_0^L h(x,0)v_k dx$. Each equation of eq (13) has a unique solution since $\mathbf{C}$ is positive definite. When $\mathbf{q}$ is not easily obtained by integration or $u(x,0)$ is only given at r discrete points, then $\boldsymbol{\alpha}(0)$ can be obtained by solving the following least-squares problem of minimizing the expression

$$\sum_{k=1}^{r}\left[u(x_k,0)-\sum_{i=1}^{2N}\alpha_i(0)v_i(x_k)\right]^2. \tag{14}$$

The value $\boldsymbol{\beta}(0)$ can be obtained in a similar manner. We shall leave the theoretical discussion here except to note that convergence of U to u or P to h has been established in the literature by Swartz and Wendroff (1969) among others. Unfortunately, the nonlinear primitive equations are not covered by Swartz and Wendroff (1969).

c. Numerical Procedure

To solve eq (10) numerically, we must first choose a basis for S_{2N}. Trigonometric functions have been widely used for such a purpose. Recently, Orszag (1970) used such functions as the basis functions when he applied the Galerkin method to the Navier-Stokes equations for incompressible flow. In recent years, spline functions have become popular in such applications. Spline functions are piecewise polynomials. For instance, Douglas et al. (1969) used cubic splines in their application of the Galerkin method to solve a petroleum engineering problem. Cubic splines have the advantage of being easily differentiable and integrable, and they usually render the matrices $\mathbf{C}$ and $\mathbf{A}$ in eq (10) very sparse. Hermite cubic functions are chosen for this study.

First, choose a partition of the region $[0, L]$; that is,

$$0=x_1<x_2<x_3<\ldots<x_N=L.$$

Then, at each node, x_i, introduce two functions

$$v_{2i-1}=\begin{cases}-2\left(\dfrac{x-x_{i-1}}{x_i-x_{i-1}}\right)^3+3\left(\dfrac{x-x_{i-1}}{x_i-x_{i-1}}\right)^2, & x\in[x_{i-1},\,x_i]\\[2ex] 2\left(\dfrac{x-x_i}{x_{i+1}-x_i}\right)^3-3\left(\dfrac{x-x_1}{x_{i+1}-x_i}\right)^2+1, & x\in[x_i,\,x_{i+1}]\\[2ex] 0, & x\in[0,\,L],\ x\notin[x_{i-1},\,x_{i+1}]\end{cases} \tag{15}$$

and

$$v_{2i}=\begin{cases}\left(\dfrac{x-x_{i-1}}{x_i-x_{i-1}}\right)^2(x-x_i), & x\in[x_{i-1},\,x_i]\\[2ex] (x-x_i)\left(\dfrac{x_{i+1}-x}{x_{i+1}-x_i}\right)^2, & x\in[x_i,\,x_{i+1}]\\[2ex] 0, & x\in[0,\,L],\ x\notin[x_{i-1},\,x_{i+1}].\end{cases} \tag{16}$$

The detailed derivation of these functions can be found in Goel (1968). We note that both functions are identically zero except in the interval (x_{i-1}, x_{i+1}). It is this property that causes the resulting matrices, $\mathbf{C}$ and $\mathbf{A}$, to be extremely sparse. The values and first derivatives of these basis functions are continuous across the nodes. We also note that v_{2i-1} has value one and first derivative zero at x_i while v_{2i} has value zero and first derivative one there. These are indeed the criteria used in deriving the functions in the first place. It is important to note that both $\mathbf{C}$ and $\mathbf{A}$ are constant matrices; therefore, they need only to be generated once at the beginning. They are very simple to generate. For instance, if we subdivide the region $(0, L)$ into three equal subregions and let $h=x_{i+1}-x_i$, then $N=4$,

$$\mathbf{C}=\frac{h^2}{420}\begin{bmatrix}\begin{pmatrix}\frac{156}{h} & 22\\ 22 & 4h\end{pmatrix} & \begin{pmatrix}\frac{54}{h} & -13\\ 13 & -3h\end{pmatrix} & & \\ \begin{pmatrix}\frac{54}{h} & 13\\ -13 & -3h\end{pmatrix} & \begin{pmatrix}\frac{312}{h} & 0\\ 0 & 8h\end{pmatrix} & \begin{pmatrix}\frac{54}{h} & -13\\ 13 & -3h\end{pmatrix} & \\ & \begin{pmatrix}\frac{54}{h} & 13\\ -13 & -3h\end{pmatrix} & \begin{pmatrix}\frac{312}{h} & 0\\ 0 & 8h\end{pmatrix} & \begin{pmatrix}\frac{54}{h} & -13\\ 13 & -3h\end{pmatrix}\\ & & \begin{pmatrix}\frac{54}{h} & 13\\ -13 & -3h\end{pmatrix} & \begin{pmatrix}\frac{156}{h} & -22\\ -22 & 4h\end{pmatrix}\end{bmatrix} \tag{17}$$

and

$$\mathbf{A}=\frac{1}{60}\begin{bmatrix}\begin{pmatrix}-30 & 6h\\ -6h & 0\end{pmatrix} & \begin{pmatrix}30 & -6h\\ 6h & -h^2\end{pmatrix} & & \\ \begin{pmatrix}-30 & -6h\\ 6h & h^2\end{pmatrix} & \begin{pmatrix}0 & 12h\\ -12h & 0\end{pmatrix} & \begin{pmatrix}30 & -6h\\ 6h & -h^2\end{pmatrix} & \\ & \begin{pmatrix}-30 & -6h\\ 6h & h^2\end{pmatrix} & \begin{pmatrix}0 & 12h\\ -12h & 0\end{pmatrix} & \begin{pmatrix}30 & -6h\\ 6h & -h^2\end{pmatrix}\\ & & \begin{pmatrix}-30 & -6h\\ 6h & h^2\end{pmatrix} & \begin{pmatrix}30 & 6h\\ -6h & 0\end{pmatrix}\end{bmatrix}. \tag{18}$$

We see that both $\mathbf{C}$ and $\mathbf{A}$ are in block tridiagonal form with each block being a 2×2 matrix. We see also that the 2×2 blocks are repeated along its diagonals except the first and last blocks on the main diagonal. Hence, both $\mathbf{C}$ and $\mathbf{A}$ are very easy to generate. Furthermore, we can easily scale the matrices such that the elements of each matrix are of the same order of magnitude simply by multiplying the basis functions v_{2i} by h^{-1}.

Any standard numerical technique can be employed to solve the systems of ordinary differential equations [eq(10)]. We chose the Crank-Nicolson method or the trapezoidal implicit method (Kurihara 1965). Let the superscript denote the time step; then the difference

equations can be written

$$\mathbf{C}(\boldsymbol{\alpha}^{n+1}-\boldsymbol{\alpha}^n)+\frac{\Delta t}{2}\,g\mathbf{A}(\boldsymbol{\beta}^{n+1}-\boldsymbol{\beta}^n)=-\Delta t g\mathbf{A}\boldsymbol{\beta}^n \quad (19a)$$

and

$$\frac{\Delta t}{2}\,H\mathbf{A}(\boldsymbol{\alpha}^{n+1}-\boldsymbol{\alpha}^n)+\mathbf{C}(\boldsymbol{\beta}^{n+1}-\boldsymbol{\beta}^n)=-\Delta t H\mathbf{A}\boldsymbol{\alpha}^n. \quad (19b)$$

We have purposely written the above equations in terms of the differences of variables at two time steps to reduce the roundoff errors. We can rearrange the equations and variables in eq (19) so that the coefficient matrix is in block tridiagonal form with each block being a 4 x 4 matrix. If we use the Gaussian elimination method to solve the final system, then the forward elimination phase need only be done once at the first time step. For each subsequent time step, one needs only to solve a decomposed system. This calculation requires only $44(N-1)$ multiplications and additions.

d. Implementation of Boundary Conditions

We have not yet discussed the means of implementing the boundary conditions. Recall that for each of the original partial differential equations there is one ordinary differential equation associated with each basis function, v_i, for $i=1,2,\ldots,2N$. To set u at the left-hand end ($x=x_1=0$), we replace the equation that comes from

$$\int_0^L (u_t+gh_x)v_1(x)dx=0 \quad (20)$$

by

$$\alpha_1(t)=u(0,t). \quad (21)$$

Likewise, to set u at the right-hand end ($x=x_N=L$), we simply replace the equation that comes from

$$\int_0^L (u_t+gh_x)v_{2N-1}dx=0 \quad (22)$$

by

$$\alpha_{2N-1}(t)=u(L,t). \quad (23)$$

If, instead, the slope of height were specified at the left end, then we would replace the equation that comes from

$$\int_0^L \left(\frac{\partial h}{\partial t}+H\frac{\partial h}{\partial x}\right)v_2dx=0 \quad (24)$$

by

$$\beta_2(x)=\frac{\partial h}{\partial x}(0,t). \quad (25)$$

Conditions (21) and (23) can be realized in system (19a) by replacing its first and next-to-last equations by

$$\alpha_1^{n+1}(t)-\alpha_1^n(t)=u(0,t)-\alpha_1^n(t) \quad (26a)$$

and

$$\alpha_{2N-1}^{n+1}(t)-\alpha_{2N-1}^n(t)=u(L,t)-\alpha_{2N-1}^n(t), \quad (26b)$$

respectively.

TABLE 1.—*Ratio of* c_τ *to* c

σ \ γ	π	$\frac{2\pi}{5}$	$\frac{\pi}{5}$	$\frac{2\pi}{15}$	$\frac{\pi}{10}$	$\frac{2\pi}{25}$	$\frac{\pi}{15}$
0.1	0.000	0.930	0.995	0.999	1.000	1.000	1.000
.2	.000	.927	.994	.998	0.999	1.000	1.000
.3	.000	.922	.992	.998	.999	0.999	1.000
.4	.000	.915	.990	.997	.998	.999	0.999
.5	.000	.906	.987	.995	.998	.999	.999
.6	.000	.895	.984	.994	.997	.998	.999
.7	.000	.884	.980	.992	.996	.997	.998
.8	.000	.871	.975	.990	.994	.997	.998
.9	.000	.857	.970	.987	.993	.996	.997
1.0	.000	.842	.964	.985	.992	.995	.996
1.1	.000	.827	.958	.982	.990	.994	.996
1.2	.000	.812	.952	.979	.988	.992	.995

4. FINITE-DIFFERENCE METHOD

The Crank-Nicolson method applied to eq (2), using fourth-order central differences to approximate the spatial derivatives, takes the form

$$u_j^{n+1}-u_j^n+\frac{g\Delta t}{2}(D_xh_j^{n+1}+D_xh_j^n)=0$$

and (27)

$$h_j^{n+1}-h_j^n+\frac{H\Delta t}{2}(D_xu_j^{n+1}+D_xu_j^n)=0$$

where n and j are the positions on the time and space coordinates, respectively, and D_x is the fourth-order central difference operator with the following effect on the operand, f_i:

$$D_xf_i=\frac{2}{3\Delta x}(f_{i+1}-f_{i-1})-\frac{1}{12\Delta x}(f_{i+2}-f_{i-2}).$$

The use of the fourth-order approximation results in eq (27) having order of accuracy comparable with the Galerkin approximation found by the cubic Hermite functions.

We shall determine the amplitude and phase characteristics of the difference equations [eq (27)]. Let u and h have the elementary solutions

$$u_j^n=u_0R^ne^{i(kn\Delta t+\gamma j\Delta x)}$$

and (28)

$$h_j^n=h_0R^ne^{i(kn\Delta t+\gamma j\Delta x)}$$

where R is the amplification factor, k is the frequency, and γ is the wave number. Substituting eq (28) into (27) and satisfying the vanishing of the determinant of the coefficients for the existence of nontrivial solutions for u_0 and h_0 results in the expression,

$$Re^{ik\Delta t}=\frac{1+i\sigma q}{1-i\sigma q}, \quad (29)$$

where $\sigma=c\Delta t/\Delta x$, $q=\frac{2}{3}\sin\gamma\Delta x-\frac{1}{12}\sin 2\gamma\Delta x$, and $c=\sqrt{gH}$ is the true phase speed. We note that $R=1$ in

eq (29). This means that eq (27) is neutral with respect to amplification of waves. The computed phase speed can be defined as

$$c_\gamma=\frac{k}{\gamma}. \tag{30}$$

From eq (29), this quantity can be expressed as

$$c_\gamma=\frac{1}{\gamma\Delta t}\arctan\frac{2\sigma q}{1-\sigma^2q^2}. \tag{31}$$

Table 1 shows the values of c_γ/c for different values of wave number, γ, and mesh ratio, σ.

5. COMPARISON OF NUMERICAL RESULTS

A number of numerical experiments were carried out (table 2) using the following initial conditions:

$$u(x, 0)=U_0\sin\frac{2\pi rx}{L}$$
and
$$h(x, 0)=H=9.184\text{ km} \tag{32}$$

where r is an integer, U_0 is the amplitude, and L=10,500 km. We further use the periodic condition given in eq (4) and (5).

Using eq (32) and periodic conditions, we obtain the exact solution to eq (2) in the form of a standing wave; that is,

$$h(x, t)=H-\frac{HU_0}{c}\cos\frac{2\pi rx}{L}\sin\frac{2\pi rct}{L}$$
and
$$u(x, t)=U_0\sin\frac{2\pi rx}{L}\cos\frac{2\pi rct}{L}. \tag{33}$$

This solution is the standard against which numerical solutions of eq (2) will be compared. Numerical computations are performed using eq (19) and (27). The use of the periodic conditions changes the form of the matrices, **C** and **A**, in eq (10). They are no longer in block-tridiagonal form. For instance, if N=4,

$$\mathbf{C}=\frac{1}{420h}\begin{bmatrix} \begin{pmatrix}312&0\\0&8\end{pmatrix} & \begin{pmatrix}54&-13\\13&-3\end{pmatrix} & & \begin{pmatrix}54&13\\-13&-3\end{pmatrix} \\ \begin{pmatrix}54&13\\-13&-3\end{pmatrix} & \begin{pmatrix}312&0\\0&8\end{pmatrix} & \begin{pmatrix}54&-13\\13&-3\end{pmatrix} & \\ & \begin{pmatrix}54&13\\-13&-3\end{pmatrix} & \begin{pmatrix}312&0\\0&8\end{pmatrix} & \begin{pmatrix}54&-13\\13&-3\end{pmatrix} \\ \begin{pmatrix}54&-13\\13&-3\end{pmatrix} & & \begin{pmatrix}54&13\\-13&-3\end{pmatrix} & \begin{pmatrix}312&0\\0&8\end{pmatrix} \end{bmatrix} \tag{34}$$

where scaling on v_{2i} by h^{-1} is assumed.

We can easily rearrange the equations and variables in eq (19) so that the final coefficient matrix is of order $4N\times 4N$ and is in the form given in eq (34) with 4×4 blocks. To solve such a system using the Gaussian elimination method, one must make approximately 76 $(N-2)$ multiplications and additions after the initial time step. The matrix resulting from finite-difference implicit equations [eq 27)] is in almost block-penta-diagonal form with 2×2 blocks. To solve such a system, one must make approximately 34 $(p-4)$ multiplications and additions for a general time step, where p is the number of spatial mesh points. The total energy of eq (2) is conserved and is given by

$$E=\frac{1}{2}\int_0^L(u^2+gh)h\,dx. \tag{35}$$

The available energy in the model is defined as

$$AE=E-\frac{1}{2}LgH^2. \tag{36}$$

The total energy for the Galerkin method, E_g, can be expressed as a function of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ by substituting eq (8) into (35). The result is the scalar product,

$$E_g=\frac{1}{2}\langle\boldsymbol{\beta},\, g\mathbf{C}\boldsymbol{\beta}+\mathbf{E}(\boldsymbol{\alpha})\boldsymbol{\alpha}\rangle, \tag{37}$$

where $\mathbf{E}(\boldsymbol{\alpha})$ is a matrix with (i, j) element

$$E_{ij}(\boldsymbol{\alpha})=\sum_{k=1}^{2N}\alpha_kQ_{kij} \tag{38}$$

and where

$$Q_{kij}=\int_0^L v_kv_iv_j\,dx. \tag{39}$$

Therefore, the available energy for the Galerkin calculation can be checked by the evaluation of the quantity $E_g-1/2\ LgH^2$. For the finite-difference calculations, the available energy is estimated by using the expression

$$\frac{1}{2}\Delta x\sum_i\left[h_iu_i^2+g(h_i-H)^2\right] \tag{40}$$

where the summation is over all the computation points.

The available energies for all the cases listed in table 2 are conserved to within 0.1 percent. To facilitate the comparison of the methods, we computed the maximum error in height at each time step for all the cases. For the finite-difference method, this error is

$$e_j^n=\max_j|h(j\Delta x, n\Delta t)-h_j^n|.$$

For the Galerkin method, the maximum error may be written

$$E^n=\max_{x\in[0,\,L]}|h(x, n\Delta t)-P(x, n\Delta t)|.$$

TABLE 2.—*Experiments performed to compare the Galerkin and finite-difference methods*

	Galerkin method No. of basis functions (time step in min)	Finite difference method No. of mesh points (time step in min)
Initial condition I	6(30, 10)	12(10, 2.5)
U_0=54.6 m/s	12(10, 2.5)	15(10, 2.5)
r=1	18(5, 1.25)	24(5, 1.25)
Initial condition II	10(10, 2.5)	15(10, 2.5)
U_0=27.3 m/s	18(5, 1.25)	30(2.5, 0.625)
r=2	24(2.5, 0.625)	45(1.25, 0.3125)

In the experiments, we estimated E^n by selecting the maximum absolute value among the errors evaluated at 200 equally spaced points. In addition, the maximum error of the derivative of height,

$$dE^n = \max_{x\in[0,\,L]} \left| \frac{\partial h}{\partial x}(x, n\Delta t) - \frac{\partial P}{\partial x}(x, n\Delta t) \right|,$$

is also estimated at each time step.

Tables 3 and 4 show the maximum errors in height normalized by the average height, H, at 5-hr intervals. Invariably, for all the cases shown, a periodic nature of e_j^n is observed with a period equal to the true period of the solution. The true periods of the solutions are 10 hr for initial condition I and 5 hr for initial condition II, respectively. The error, e_j^n, has a relative maximum at each half period. We can explain the periodic nature of e_j^n as follows: the main error in these calculations is the truncation error in time, and this error is proportional to the third derivative, $\partial^3 h/\partial t^3$, for a method correct to second order.

The behavior of E^n has a similar periodic character when the time step used is large, such that the truncation error in time is dominant. However, when the time step is reduced, the error due to initial projection plus the roundoff error become important. As a consequence of this, E^n behaves randomly as shown by column 5 in table 5. In any case, the errors listed in tables 5–8 are the maximums encountered during the 5-hr time segment ending at the particular hours indicated in the left column. Tables 7–8 show the error, dE^n, normalized by the maximum value of $\partial h/\partial x$.

It is obvious from these experiments that the Galerkin approximations are much more accurate than the finite-

TABLE 3.—*Errors using the finite-difference method,* e_j^n/H, *for initial condition I*

Hr	12 mesh points		15 mesh points		24 mesh points	
	time step (min) 10	2.5	time step (min) 10	2.5	time step (min) 5	1.25
5	1.84D−3*	1.37D−3	1.09D−3	6.07D−4	2.17D−4	9.56D−5
10	3.68D−3	2.74D−3	2.19D−3	1.21D−3	4.34D−4	1.91D−4
15	5.52D−3	4.11D−3	4.28D−3	1.82D−3	6.57D−4	2.87D−4
20	7.36D−3	5.48D−3	4.38D−3	2.43D−3	8.68D−4	3.82D−4
25	9.19D−3	6.85D−3	5.47D−3	3.03D−3	1.08D−3	4.78D−4
30	1.10D−2	8.22D−3	6.56D−3	3.64D−3	1.30D−3	5.73D−4
35	1.29D−2	9.59D−3	7.66D−3	4.25D−3	1.52D−3	6.69D−4
40	1.47D−2	1.10D−2	8.75D−3	4.85D−3	1.74D−3	7.64D−4
45	1.65D−2	1.23D−2	9.85D−3	5.46D−3	1.95D−3	8.60D−4
50	1.84D−2	1.37D−2	1.09D−2	6.07D−3	2.17D−3	9.55D−4
55	2.02D−2	1.51D−2	1.20D−2	6.67D−3	2.39D−3	1.05D−3
60	2.20D−2	1.64D−2	1.31D−2	7.28D−3	2.60D−3	1.15D−3

*1.84D−3=1.84×10⁻³.

TABLE 4.—*Errors using the finite-difference method,* e_j^n/H, *for initial condition II*

Hr	15 mesh points		30 mesh points		45 mesh points	
	time step (min) 10	2.5	time step (min) 2.5	0.625	time step (min) 1.25	0.3125
5	1.06D−2	8.75D−3	7.04D−4	5.82D−4	1.47D−4	1.16D−4
10	2.11D−2	1.74D−2	1.42D−3	1.16D−3	2.94D−4	2.33D−4
15	3.12D−2	2.59D−2	2.11D−3	1.75D−3	4.41D−4	3.49D−4
20	4.10D−2	3.42D−2	2.82D−3	2.33D−3	5.87D−4	4.65D−4
25	5.03D−2	4.22D−2	3.52D−3	2.91D−3	7.34D−4	5.81D−4
30	5.89D−2	4.99D−2	4.22D−3	3.49D−3	8.81D−4	6.97D−4
35	6.67D−2	5.70D−2	4.93D−3	4.07D−3	1.03D−3	8.14D−4
40	7.37D−2	6.37D−2	5.63D−3	4.66D−3	1.17D−3	9.30D−4
45	7.97D−2	6.99D−2	6.34D−3	5.24D−3	1.32D−3	1.05D−3
50	8.47D−2	7.54D−2	7.04D−3	5.82D−3	1.47D−3	1.16D−3
55	8.86D−2	8.03D−2	7.74D−3	6.40D−3	1.62D−3	1.28D−3
60	9.14D−2	8.45D−2	8.45D−3	6.98D−3	1.76D−3	1.39D−3

TABLE 5.—*Errors in Galerkin approximation,* E^n/H, *for initial condition I*

Hr	6 basis functions		12 basis functions		18 basis functions	
	time step (min) 30	10	time step (min) 10	2.5	time step (min) 5	1.25
5	4.61D−3	1.75D−3	5.50D−4	2.91D−4	1.48D−4	8.06D−5
10	9.07D−3	1.93D−3	9.85D−4	2.13D−4	2.71D−4	9.96D−5
15	1.38D−2	2.00D−3	1.61D−3	2.50D−4	3.99D−4	7.32D−5
20	1.81D−2	2.10D−3	2.04D−3	2.81D−4	5.35D−4	1.01D−4
25	2.30D−2	2.55D−3	2.60D−3	2.65D−4	6.50D−4	6.65D−5
30	2.71D−2	2.71D−3	3.14D−3	2.60D−4	7.96D−4	6.94D−5
35	3.20D−2	3.22D−3	3.58D−3	2.49D−4	9.22D−4	9.05D−5
40	3.62D−2	3.43D−3	4.20D−3	2.92D−4	1.05D−3	9.06D−5
45	4.09D−2	3.90D−3	4.62D−3	3.10D−4	1.19D−3	9.65D−5
50	4.52D−2	4.29D−3	5.20D−3	3.20D−4	1.30D−3	8.73D−5
55	4.98D−2	4.77D−3	5.72D−3	3.63D−4	1.45D−3	9.83D−5
60	5.40D−2	5.14D−3	6.18D−3	3.59D−4	1.57D−3	9.44D−5

TABLE 6.—*Errors in Galerkin approximation,* E^n/H, *for initial condition II*

Hr	10 basis functions		18 basis functions		24 basis functions	
	time step (min) 10	2.5	time step (min) 5	1.25	time step (min) 2.5	0.625
5	1.89D−3	1.67D−3	5.67D−4	2.83D−4	1.44D−4	8.54D−5
10	3.73D−3	1.64D−3	1.02D−3	2.79D−4	2.52D−4	1.39D−4
15	5.64D−3	1.69D−3	1.57D−3	3.02D−4	4.09D−4	1.34D−4
20	7.47D−3	1.69D−3	2.04D−3	2.48D−4	5.17D−4	9.16D−5
25	9.40D−3	1.73D−3	2.60D−3	3.05D−4	6.11D−4	1.09D−4
30	1.12D−2	1.72D−3	3.06D−3	2.59D−4	7.54D−4	1.43D−4
35	1.32D−2	1.76D−3	3.60D−3	2.95D−4	9.17D−4	1.30D−4
40	1.49D−2	1.75D−3	4.07D−3	2.82D−4	1.03D−3	8.81D−5
45	1.69D−2	1.80D−3	4.64D−3	3.05D−4	1.12D−3	9.21D−5
50	1.86D−2	1.79D−3	5.10D−3	3.04D−4	1.25D−3	1.46D−4
55	2.06D−2	1.83D−3	5.63D−3	2.80D−4	1.43D−3	1.23D−4
60	2.23D−2	1.82D−3	6.10D−3	3.17D−4	1.55D−3	1.04D−4

TABLE 7.—*Errors in Galerkin approximation,* $\dfrac{dE^n}{|max\ \partial h/\partial x|}$, *for initial condition I*

Hr	6 basis functions		12 basis functions		18 basis functions	
	time step (min) 30	10	time step (min) 10	2.5	time step (min) 5	1.25
5	6.32D−2	6.14D−2	8.96D−3	1.79D−2	5.78D−3	5.97D−3
10	7.28D−2	5.76D−2	1.70D−2	1.39D−2	7.02D−3	7.21D−3
15	8.64D−2	6.08D−2	1.72D−2	1.60D−2	5.28D−3	5.51D−3
20	1.08D−1	6.34D−2	1.83D−2	1.73D−2	8.06D−3	7.39D−3
25	1.22D−1	6.24D−2	2.08D−2	1.67D−2	7.10D−3	4.84D−3
30	1.48D−1	6.43D−2	2.28D−2	1.61D−2	8.31D−3	6.82D−3
35	1.72D−1	6.40D−2	2.86D−2	1.45D−2	9.07D−3	6.54D−3
40	1.97D−1	6.52D−2	2.77D−2	1.76D−2	6.90D−3	6.62D−3
45	2.21D−1	6.56D−2	3.06D−2	1.82D−2	9.14D−3	7.03D−3
50	2.45D−1	6.61D−2	3.16D−2	1.34D−2	8.42D−3	4.67D−3
55	2.69D−1	6.72D−2	3.27D−2	1.58D−2	1.10D−2	7.22D−3
60	2.93D−1	6.70D−2	4.01D−2	1.78D−2	1.21D−2	5.48D−3

TABLE 8.—*Errors in Galerkin approximat on,* $\frac{dE^n}{|max\ \partial h/\partial x|}$, *for initial condition II*

Hr	10 basis functions		18 basis functions		24 basis functions	
	Time step (min) 10	2.5	Time step (min) 5	1.25	Time step (min) 2.5	0.625
5	1.18 $D-1$	1.16 $D-1$	2.76 $D-2$	2.68 $D-2$	1.07 $D-2$	1.07 $D-2$
10	1.27 $D-1$	1.15 $D-1$	3.04 $D-2$	2.64 $D-2$	1.75 $D-2$	1.72 $D-2$
15	1.30 $D-1$	1.16 $D-1$	2.60 $D-2$	2.79 $D-2$	1.79 $D-2$	1.68 $D-2$
20	1.38 $D-1$	1.14 $D-1$	3.12 $D-2$	2.38 $D-2$	1.45 $D-2$	1.01 $D-2$
25	1.45 $D-1$	1.17 $D-1$	3.87 $D-2$	2.84 $D-2$	1.16 $D-2$	1.13 $D-2$
30	1.62 $D-1$	1.15 $D-1$	3.78 $D-2$	2.47 $D-2$	1.83 $D-2$	1.76 $D-2$
35	1.78 $D-1$	1.18 $D-1$	4.35 $D-2$	2.74 $D-2$	2.04 $D-2$	1.63 $D-2$
40	1.94 $D-1$	1.15 $D-1$	4.89 $D-2$	2.59 $D-2$	2.00 $D-2$	9.64 $D-3$
45	2.10 $D-1$	1.19 $D-1$	5.67 $D-2$	2.78 $D-2$	1.48 $D-2$	1.19 $D-2$
50	2.25 $D-1$	1.16 $D-1$	5.60 $D-2$	2.70 $D-2$	1.90 $D-2$	1.78 $D-2$
55	2.40 $D-1$	1.19 $D-1$	6.58 $D-2$	2.50 $D-2$	2.27 $D-2$	1.58 $D-2$
60	2.59 $D-1$	1.17 $D-1$	6.79 $D-2$	2.83 $D-2$	1.51 $D-2$	9.23 $D-3$

difference results for the same amount of computing effort. If we adopt Grammeltvedt's (1969) criterion that 15 mesh points per wavelength are required to describe adequately the wave of interest, then a comparable accuracy can be obtained by only six basis functions with the Galerkin method. We also notice that reduction of time step has a greater effect on the Galerkin method than on the finite-difference method.

It appears that, to achieve a required accuracy, one should use a minimum number of basis functions combined with a time step small enough such that the round-off error just starts to dominate.

6. GALERKIN METHOD APPLIED TO NONLINEAR EQUATIONS

The Galerkin equations for the nonlinear wave equations [eq (1)] are (for $j=1, 2, \ldots, 2N$)

$$\sum_{i=1}^{2N}\int_0^L \frac{d\alpha_i}{dt} v_i v_j dx+\sum_{i=1}^{2N}\int_0^L \sum_{k=1}^{2N} \alpha_k v_k \alpha_i \frac{dv_i}{dx} v_j dx +g\sum_{i=1}^{2N}\int_0^L \beta_i \frac{dv_i}{dx} v_j dx=0 \quad (41a)$$

and

$$\sum_{i=1}^{2N}\int_0^L \frac{d\beta_i}{dt} v_i v_j dx+\sum_{i=1}^{2N}\int_0^L \sum_{k=1}^{2N} \alpha_k v_k \beta_i \frac{dv_i}{dx} v_j dx +\sum_{i=1}^{2N}\int_0^L \sum_{k=1}^{2N} \beta_k v_k \alpha_i \frac{dv_i}{dx} v_j dx=0. \quad (41b)$$

If we let

$$w_{kij}=\int_0^L v_k \frac{dv_i}{dx} v_j dx, \quad (42)$$

then eq (41) may be written in the matrix notation

$$\mathbf{C}\frac{d\boldsymbol{\alpha}}{dt}+\mathbf{B}(\boldsymbol{\alpha})\boldsymbol{\alpha}+g\mathbf{A}\boldsymbol{\beta}=0$$

and

$$\mathbf{C}\frac{d\boldsymbol{\beta}}{dt}+\mathbf{B}(\boldsymbol{\alpha})\boldsymbol{\beta}+\mathbf{B}(\boldsymbol{\beta})\boldsymbol{\alpha}=0 \quad (43)$$

where matrices $\mathbf{C}$ and $\mathbf{A}$ are the same as those defined in the linear case; and where $\mathbf{B}(\boldsymbol{\alpha})$ and $\mathbf{B}(\boldsymbol{\beta})$ are matrices for which the (i, j) elements are

$$B_{ij}(\boldsymbol{\alpha})=\sum_{k=1}^{2N}\alpha_k w_{kij}$$

and

$$B_{ij}(\boldsymbol{\beta})=\sum_{k=1}^{2N}\beta_k w_{kij}. \quad (44)$$

Again, the Galerkin method has transformed the original partial differential equations into a system of ordinary differential equations. Since the matrices $\mathbf{B}(\boldsymbol{\alpha})$ and $\mathbf{B}(\boldsymbol{\beta})$ depend on $\boldsymbol{\alpha}$ or $\boldsymbol{\beta}$, eq (43) are nonlinear. Using the Crank-Nicolson method in discretizing the time domain and writing the equations again in terms of the differences of variables at two time steps, we have

$$\left[\mathbf{C}+\frac{\Delta t}{2}\mathbf{B}\left(\frac{\boldsymbol{\alpha}^{n+1}+\boldsymbol{\alpha}^n}{2}\right)\right](\boldsymbol{\alpha}^{n+1}-\boldsymbol{\alpha}^n)+\frac{\Delta t}{2}g\mathbf{A}(\boldsymbol{\beta}^{n+1}-\boldsymbol{\beta}^n) =-\Delta t\mathbf{B}\left(\frac{\boldsymbol{\alpha}^{n+1}+\boldsymbol{\alpha}^n}{2}\right)\boldsymbol{\alpha}^n-\Delta t g\mathbf{A}\boldsymbol{\beta}^n$$

and

$$\frac{\Delta t}{2}\mathbf{B}\left(\frac{\boldsymbol{\beta}^{n+1}+\boldsymbol{\beta}^n}{2}\right)(\boldsymbol{\alpha}^{n+1}-\boldsymbol{\alpha}^n) +\left[\mathbf{C}+\frac{\Delta t}{2}\mathbf{B}\left(\frac{\boldsymbol{\alpha}^{n+1}+\boldsymbol{\alpha}^n}{2}\right)\right](\boldsymbol{\beta}^{n+1}-\boldsymbol{\beta}^n) =-\Delta t\mathbf{B}\left(\frac{\boldsymbol{\beta}^{n+1}+\boldsymbol{\beta}^n}{2}\right)\boldsymbol{\alpha}^n-\Delta t\mathbf{B}\left(\frac{\boldsymbol{\alpha}^{n+1}+\boldsymbol{\alpha}^n}{2}\right)\boldsymbol{\beta}^n. \quad (45)$$

Equation (45) is a system of nonlinear algebraic equations.

If we use Hermite cubic functions as basis functions, the matrix $\mathbf{B}$ again has the block tridiagonal form with 2×2 blocks the same as the form of matrices $\mathbf{C}$ and $\mathbf{A}$. But, unlike $\mathbf{C}$ and $\mathbf{A}$, $\mathbf{B}$ needs to be recomputed at each time step because of its dependence on $\boldsymbol{\alpha}$ or $\boldsymbol{\beta}$. The computation of $\mathbf{B}$ is also more complicated than the computation of $\mathbf{C}$ and $\mathbf{A}$ since integrals of triple products of basis functions and their derivatives [eq (42)] need to be evaluated first. Then, an inner product [eq (44)] has to be computed for each nonzero entry of $\mathbf{B}$. Fortunately, we do not have too many integrals of triple products to evaluate; most of them are zero since each basis function has a nonzero value only in two neighboring subregions. In fact, for any node, x_m, eq (42) may be written as

$$w_{kij}=\int_{-\Delta x_{m-1}}^0 v_k\frac{dv_i}{dx}v_j dx+\int_0^{\Delta x_m} v_k\frac{dv_i}{dx}v_j dx \quad (46)$$

where we note from figure 1 that the maximum difference of any two indices of the triple product is three. There are exactly 40 different integrals to evaluate for each term on the right side of eq (46). These integrals need only be evaluated once. On the other hand, eq (44) has to be evaluated repeatedly every time step for every nonzero element of $\mathbf{B}$. Since, however, only six terms enter into the inner product [eq (44)], evaluation of $\mathbf{B}$ is not a major computational effort.

To avoid solving nonlinear algebraic systems of equations, we use the following predictor-corrector approxi-

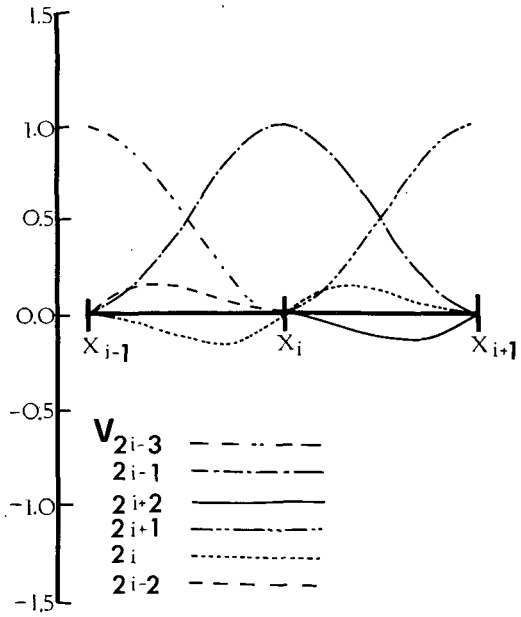

FIGURE 1.—Hermite cubic basis functions for the subregion (x_{i-1}, x_{i+1}), where the node spacing is unity.

mation to eq (45) (Douglas and Dupont 1970):

Predictor

$$\left[\mathbf{C}+\frac{\Delta t}{2}\mathbf{B}(\boldsymbol{\alpha}^n)\right](\bar{\boldsymbol{\alpha}}^{n+1}-\boldsymbol{\alpha}^n)+\frac{\Delta t}{2}g\mathbf{A}(\bar{\boldsymbol{\beta}}^{n+1}-\boldsymbol{\beta}^n)$$
$$=-\Delta t\mathbf{B}(\boldsymbol{\alpha}^n)\boldsymbol{\alpha}^n-\Delta t g\mathbf{A}\boldsymbol{\beta}^n$$

and (47a)

$$\frac{\Delta t}{2}\mathbf{B}(\boldsymbol{\beta}^n)(\bar{\boldsymbol{\alpha}}^{n+1}-\boldsymbol{\alpha}^n)+\left[\mathbf{C}+\frac{\Delta t}{2}\mathbf{B}(\boldsymbol{\alpha}^n)\right](\bar{\boldsymbol{\beta}}^{n+1}_{+}-\boldsymbol{\beta}^n)$$
$$=-\Delta t\mathbf{B}(\boldsymbol{\beta}^n)\boldsymbol{\alpha}^n-\Delta t\mathbf{B}(\boldsymbol{\alpha}^n)\boldsymbol{\beta}^n.$$

Corrector

$$\left[\mathbf{C}+\frac{\Delta t}{2}\mathbf{B}\left(\frac{\bar{\boldsymbol{\alpha}}^{n+1}+\boldsymbol{\alpha}^n}{2}\right)\right](\boldsymbol{\alpha}^{n+1}-\boldsymbol{\alpha}^n)+\frac{\Delta t}{2}g\,\mathbf{A}\,(\boldsymbol{\beta}^{n+1}-\boldsymbol{\beta}^n)$$
$$=-\Delta t\;\mathbf{B}\left(\frac{\bar{\boldsymbol{\alpha}}^{n+1}+\boldsymbol{\alpha}^n}{2}\right)\boldsymbol{\alpha}^n-\Delta t g\;\mathbf{A}\,\boldsymbol{\beta}^n$$

and (47b)

$$\frac{\Delta t}{2}\mathbf{B}\left(\frac{\bar{\boldsymbol{\beta}}^{n+1}+\boldsymbol{\beta}^n}{2}\right)(\boldsymbol{\alpha}^{n+1}-\boldsymbol{\alpha}^n)$$
$$+\left[\mathbf{C}+\frac{\Delta t}{2}\mathbf{B}\left(\frac{\bar{\boldsymbol{\alpha}}^{n+1}+\boldsymbol{\alpha}^n}{2}\right)\right](\boldsymbol{\beta}^{n+1}-\boldsymbol{\beta}^n)$$
$$=-\Delta t\,\mathbf{B}\left(\frac{\bar{\boldsymbol{\beta}}^{n+1}+\boldsymbol{\beta}^n}{2}\right)\boldsymbol{\alpha}^n-\Delta t\;\mathbf{B}\left(\frac{\bar{\boldsymbol{\alpha}}^{n+1}-\boldsymbol{\alpha}^n}{2}\right)\boldsymbol{\beta}^n.$$

We note that the predictor-corrector requires the solution of two sparse $4N\times 4N$ linear algebraic systems for each time step.

Douglas and Dupont (1970) also considered another procedure that requires the solution of only one system of algebraic equations. They called this procedure Crank-Nicolson extrapolation. When applied to eq(45), they become

$$\left[\mathbf{C}+\frac{\Delta t}{2}\mathbf{B}\left(\frac{3}{2}\boldsymbol{\alpha}^n-\frac{1}{2}\boldsymbol{\alpha}^{n-1}\right)\right](\boldsymbol{\alpha}^{n+1}-\boldsymbol{\alpha}^n)+\frac{\Delta t}{2}g\mathbf{A}(\boldsymbol{\beta}^{n+1}-\boldsymbol{\beta}^n)$$
$$=-\Delta t\mathbf{B}\left(\frac{3}{2}\boldsymbol{\alpha}^n-\frac{1}{2}\boldsymbol{\alpha}^{n-1}\right)\boldsymbol{\alpha}^n-\Delta t g\mathbf{A}\boldsymbol{\beta}^n$$

and (48)

$$\frac{\Delta t}{2}\mathbf{B}\left(\frac{3}{2}\boldsymbol{\beta}^n-\frac{1}{2}\boldsymbol{\beta}^{n-1}\right)(\boldsymbol{\alpha}^{n+1}-\boldsymbol{\alpha}^n)$$
$$+\left[\mathbf{C}+\frac{\Delta t}{2}\mathbf{B}\left(\frac{3}{2}\boldsymbol{\alpha}^n-\frac{1}{2}\boldsymbol{\alpha}^{n-1}\right)\right](\boldsymbol{\beta}^{n+1}-\boldsymbol{\beta}^n)$$
$$=-\Delta t\mathbf{B}\left(\frac{3}{2}\boldsymbol{\beta}^n-\frac{1}{2}\boldsymbol{\beta}^{n-1}\right)\boldsymbol{\alpha}^n-\Delta t\mathbf{B}\left(\frac{3}{2}\boldsymbol{\alpha}^n-\frac{1}{2}\boldsymbol{\alpha}^{n-1}\right)\boldsymbol{\beta}^n.$$

This extrapolation scheme requires a special starting procedure since it involves values on three time levels. Douglas and Dupont (1970) proved that for parabolic equations both the predictor-corrector and the extrapolation schemes retain the second-order accuracy in time of the original Crank-Nicolson-Galerkin method.

7. NONLINEAR NUMERICAL RESULTS

In the numerical example given below, we again use the periodic condition for comparison with the finite-difference results free from boundary effects. For the Galerkin calculation, we use the predictor-corrector formula [eq (47)] for the first time step and the extrapolated formula [eq (48)] for subsequent time steps. For the finite-difference calculation, we use the predictor-corrector formula of Gourlay and Morris (1968) except that fourth-order finite differences are used to approximate $\partial u/\partial x$ and $\partial h/\partial x$. These are:

Predictor

$$u_i^{*n+1}=\frac{1}{2}(u_{i+1}^n+u_{i-1}^n)-\Delta t(u_i^n D_x u_i^n+gD_x h_i^n)$$

and (49a)

$$h_i^{*n+1}=\frac{1}{2}(h_{i+1}^n+h_{i-1}^n)-\Delta t(h_i^n D_x u_i^n+u_i^n D_x h_i^n).$$

Corrector

$$u_i^{n+1}=u_i^n-\frac{\Delta t}{2}\left[u_i^n D_x u_i^n+gD_x h_i^n+u_i^{*n+1}D_x u_i^{n+1}+gD_x h_i^{n+1}\right]$$

and (49b)

$$h_i^{n+1}=h_i^n-\frac{\Delta t}{2}\left[h_i^n D_x u_i^n+u_i^n D_x h_i^n+h_i^{*n+1}D_x u_i^{n+1}+u_i^{*n+1}D_x h_i^{n+1}\right].$$

Using Gaussian elimination, the operation count is 480 $(N-2)$ multiplications-additions for eq (48) disregarding the evaluations of the **B** matrix. In the case of the finite-difference formulas [eq (49)], the count is $198(p-4)$ multiplications-additions for the corrector alone. Numeri-

TABLE 9.—*Maximum deviation from the standard solution*

Hr	Galerkin solution			Finite-difference solution		
	No. of basis functions time step (min)			No. of mesh points time step (min)		
	6(15)	12(5)	18(2.5)	9(15)	15(5)	25(2.5)
10	1.04D-3	1.03D-4	3.75D-5	4.99D-3	1.15D-3	2.14D-4
20	3.06D-3	5.22D-4	1.52D-4	1.06D-2	4.16D-3	1.37D-3
30	7.01D-3	2.56D-3	6.93D-4	1.88D-2	7.85D-3	4.09D-3
40	1.58D-2	5.95D-3	1.80D-3	3.06D-2	1.66D-2	7.78D-3
50	2.64D-2	7.70D-3	3.24D-3	3.88D-2	2.75D-2	1.18D-2

cal experiments were carried out using both methods for initial conditions $u(x, 0)=13.65 \sin(2\pi x/L)$ m/s and $h(x, 0)=9184$ m. Again, both methods were able to preserve the available energy very well. Results from these experiments are compared with the "standard" solution obtained by using 36 basis functions and a time step of 0.625 min. Table 9 shows the maximum deviations from the standard height solution at $L/2$ normalized by the initial height for each 10-hr period. We note that, for comparable amounts of computation, the Galerkin results are in closer agreement with the standard solution than are the corresponding finite-difference solutions.

8. CONCLUSIONS

We have demonstrated that the Galerkin method combined with piecewise, cubic, Hermite basis functions can be efficiently applied to both linear and nonlinear wave equations. We have shown experimentally that, for the same accuracy requirement, the Galerkin procedure needs less computation than the corresponding finite-difference method. This, in turn, means less computer storage for the Galerkin method. In addition, the Galerkin procedure presented here produces smooth, global approximations to the solutions of eq (1) and (2) and their first spatial derivatives.

It is proper to mention that eq (10) and (43) apply equally well to nonuniform node spacings. The generation of coefficient matrices **A**, **B**, and **C** and the solution of eq (19) and (47) would present no more difficulty in the nonuniform case than in the uniform case. Such computations are currently under investigation.

The major disadvantage of the Galerkin procedure is that it is more difficult to program for a computer. However, in view of the advantages listed above, we feel that it is a viable alternative to finite-difference schemes.

REFERENCES

Ahlberg, J. H., Nilson, E. N., and Walsh, J. L., *The Theory of Splines and Their Applications*, Academic Press, New York, N.Y., **1967**, 284 pp.

Birkhoff, G., Schultz, M. H., and Varga, R. S., "Piecewise Hermite Interpolation in One and Two Variables With Applications to Partial Differential Equations," *Numerische Mathematik*, Vol. 11, Springer-Verlag, New York, N.Y., **1968**, pp. 232–256.

Douglas, J., and Dupont, T., "Galerkin Methods for Parabolic Problems," *SIAM Journal on Numerical Analysis*, Vol. 7, No. 4, Society for Industrial and Applied Mathematics, Philadelphia, Pa., Dec. **1970**, pp. 575–626.

Douglas J., Dupont, T., and Rachford, H. H., "The Application of Variational Methods to Waterflooding Problems," *Journal of Canadian Petroleum Technology*, Vol. 8, No. 3, Petroleum Society of Canadian Institute of Mining and Metallurgy, Montreal, Canada, July–Sept. **1969**, pp. 1–7.

Galerkin, B. G., "Rods and Plates. Series Occurring in Various Questions Concerning the Elastic Equilibrium of Rods and Plates," *Engineers Bulletin*, Vol. 19, **1915**, pp. 897–908.

Goel, J. J., "Construction of Basic Functions for Numerical Utilization of Ritz's Method," *Numerische Mathematik*, Vol. 12, Springer-Verlag, New York, N.Y., **1968**, pp. 435–447.

Gourlay, A. R., and Morris, J. Ll., "Finite-Difference Methods for Nonlinear Hyperbolic Systems," *Mathematics of Computation*, Vol. 22, No. 101, American Mathematical Society, Providence, R.I., **1968**, pp. 28–39.

Grammeltvedt, Arne, "A Survey of Finite-Difference Schemes for the Primitive Equations for a Barotropic Fluid," *Monthly Weather Review*, Vol. 97, No. 5, May **1969**, pp. 384–404.

Kurihara, Yoshio, "On the Use of Implicit and Iterative Methods for the Time Integration of the Wave Equation," *Monthly Weather Review*, Vol. 93, No. 1, Jan. **1965**, pp. 33–46.

Mikhlin, S. G., *Variational Methods in Mathematical Physics*, Pergamon Press, New York, N.Y., **1964**, 582 pp. (translation of *Variatsionnyye Metody v Matematicheskoi fizike*, Gostekhizdat, Moscow, U.S.S.R., 1957).

Moretti, G., "Importance of Boundary Conditions in the Numerical Treatment of Hyperbolic Equations," Symposium on High-Speed Computing in Fluid Dynamics, August 19–24, 1968, Monterey, Calif., *The Physics of Fluids Supplement II*, American Institute of Physics, New York, N.Y., **1969**, 291 pp. (see pp. 13–20).

Orszag, Steven A., "Accuracy of Numerical Simulation of Incompressible Flows," *NCAR Manuscript* No. 70-73, National Center for Atmospheric Research, Boulder, Colo., **1970**, 43 pp. plus diagrams.

Price, H. S., and Varga, R. S., "Error Bounds for Semidiscrete Galerkin Approximations of Parabolic Problems With Applications to Petroleum Reservoir Mechanics," *Numerical Solution of Field Problems in Continuum Physics*, American Mathematical Society, Providence, R.I., **1970**, 280 pp. (see pp. 74–94).

Swartz, B. K., and Wendroff, B., "Generalized Finite Difference Schemes," *Mathematics of Computation*, Vol. 23, No. 105, American Mathematical Society, Providence, R.I., **1969**, pp. 37–49.

Varga, Richard S., "Accurate Numerical Methods for Nonlinear Boundary Value Problems," *Numerical Solution of Field Problems in Continuum Physics*, American Mathematical Society, Providence R.I., **1970**, 280 pp. (see pp. 152–167).

[*Received October 20, 1971; revised June 5, 1972*]